International Journal of Scholarly Research (IJSR)
Vol-1, Issue-1, 2017
http://www.scholars.org.in

“Review On An approach for Mining of Datasets using Apriori Hybrid Algorithm"

Kajal R. Thakre [1], Prof.RanjanaShende [2]

[1] Departement of computer science and engg,
G.H. Raisoni Institute of Engineering and Technology,Nagpur
kajalthakre07@gmail.com

[2]Departement of computer science and engg,
G.H. Raisoni Institute of Engineering and Technology,Nagpur
ranjana.shende@raisoni.net

Abstract: Usually, data mining sometimes called data or knowledge detection is the process of analyzing data from different viewpoints and arrange it into useful information that can be used to increase profits, cuts costs, or both. Basically data mining is one kind of database where we can store large amount of data and the required data is extracted as per user need. So Data mining software is one of an analytical or data extraction which is called mining tools for examining data. As per the rapidly increasing popularity of Data mining in different firms and organization for example banking, medicine, scientific research and among government agencies there are many possible methods are available for data analysis. It allows users to analyse data from many different dimensions or angles, group it, and encapsulate the relationships recognised. It would be having the dataset of one of the store and that datasets are in particular format such as json,csv,xml . Data stored in text databases is commonly semi-structured, i.e., it is neither completely unstructured nor completely structured. For example, a document may contain a few structured fields, such as title, authors, publication date, length, and category, and so on, but also contain some largely unstructured text components, such as abstract and contents. In our paper we have done some recent database studies to design and implement semi-structured data. Data extraction procedures, such as text indexing, have been established to handle the unstructured documents such as abstract and content from above example. So in our paper we are proceeding datasets through apriori hybridalgorithm which would be a combination of weighted apriori and hash T for search result. After this extraction the search result is further proceed to the FDM(fast distributed mining algorithm) for comparison.

Keyword: Data mining, frequent item sets, apriori hybrid algorithm, FDM(fast distributed mining algorithm), java parser.

1. INTRODUCTION

As per the rapidly increasing popularity and need of Data mining in different firms and organization for example banking, medicine, scientific research and among government agencies there are many possible methods are available for data analysis. These days, in each and every organization or in firms there is requirement to store data or information which is in a large quantity. Large quantity of data is being collected in the data warehouse. Usually there is a huge gap from the stored data to the knowledge that could be constructed from the data. Because data which is stored in data warehouse is in large quantity and searching required data from this huge container is a challenging task. This conversion won't occur spontaneously, that's where Data Mining comes into picture. In Experimental Data Analysis, some initial knowledge is known about the data means we have a partial information about particular data, but Data Mining could help in a more in-depth knowledge about the data by exploring the data in detail and its similar information is also get by data mining. Pursuing knowledge from huge quantity of data is one of the most preferred features of Data Mining. Manual data analysis has been around for some time now, but it creates a bottleneck for large data analysis that requires a more manpower.

As increasing demand of data mining in the real world environment three will be a fast rising computer science and engineering techniques and methodology produces new demand to mine complex and huge amount of data types. A number of Data Mining techniques such as association, clustering, classification are come in front of us to mine this vast amount of data. This techniques have their own benefits and features. As we can describe one by one characteristics of this techniques,

1. Association rule mining:

Essentially anyone wishing to do similarity analysis on products,whether at a corporal store or at an online e-commerce store, will evaluate the use of association algorithms. The aim of association rule discovery is to find associations among items from a set of relations, each of which contains a set of items as per called it as dataset.

Not all of the association rules discovered within a transaction set are exciting or beneficial. Usually the algorithm finds a subset of association rules that fulfil certain constrictions.

2. Clustering:
   In this technique of data mining all the related items in a datasets are cluster together that means this technique gather or group many items which are close to each other. This form of mining technique is known as clustering.
3. Classification:
   Classification is a technique in which the data is extracted by passing query and as per requirements of user query the data is extracted or mines.

Previous studies on Data Mining focus on structured data, such as relational and transactional data. However, in certainty, a significant portion of the available information is stored in text databases (or document databases), which consists of large collections of documents in the form of files and folders from various sources such as news articles, books, digital libraries and Web pages. As we all known that the information knowledge of today's world is increased through the internet so collection of data in large volume is possible easily in this internet world.

Data stored in form of text is known as Text databases are quickly budding due to the increasing amount of information available in electronic forms such as electronic publications, e-mail, CD-ROMs, and the World Wide Web which can also be viewed as a huge, interconnected, dynamic text database. Data stored in text databases is mostly semi-structured, i.e., it is neither completely unstructured nor completely structured. That means in past decades data stored in fully structured form means in the form of table the database data is available. But in semi-structured database data is neither in full structured database nor completely in un-structured in form of normal text. For example in book store of books, a document may contain a limited structured fields, such as title, authors, publication date, length, category, and so on, but also contain some largely unstructured text components, such as abstract and contents which have large information and cannot be defined by one item of data so it is called unstructured. In current database research, studies have been done to model and implement semi structured data that can be defined or represented by one data item in a large datasets. Information Retrieval techniques, such as text indexing, have been developed to handle the unstructured documents.

But, old-style Information Retrieval techniques become insufficient for the rapid increasingly vast amount of text data. Typically, only a small segment of the many available documents will be appropriate to a given individual or user. Without knowing what could be in the documents, it is difficult to express effective queries for analyzing and extracting useful information from the data. Users need tools to associate different documents, rank the importance and significance of the documents, or find patterns and trends across multiple documents. Thus, Text Mining has become an increasingly popular and essential theme in Data Mining. This text mining save a lot of effort of individuals. In this paper we were proposing java parser to perform the extraction on the large data sets from that extraction we got the estimated result of datasets that we want a required and then further proceed toward APRIORI-Hybrid algorithm which is formed by combining weighted apriori and THashapriori. After this this result obtained by algorithm is transferred to FDM association rule mining algorithm for comparison of efficiency, frequency, memory consumption and other parameters.

2. RELATED WORK

As we know about data mining has getting very importance in together world so there is some different association rule applied by many other people so we have just studied there work and get some important techniques and methods that are described in detail as follows:

A TamirTassa [1] has **P**rivacy preserving data mining word done considered two related settings. One, in which the data owner and the data miner are two different entities, and another, in which the data is distributed among several parties who aim to jointly perform data mining on the unified corpus of data that they hold.He proposed a procedure for secure mining of association rules in horizontally distributed databases that improves expressively upon the current leading protocol in terms of confidentiality and effectiveness. He proposed a main ingredient as a new protocol a novel secure multi-party protocol for calculating the union (or intersection) of isolated subsets that each of the cooperating players hold. Another ingredient is a protocol that experiments the insertion of an element held by one player in a subset held by another. The latter protocol achievements the fact that the underlying problem is of interest only when the number of players is greater than two.

Merry KP, Rabindra Kumar Singh, Swaroop.S.Kumar [2] has tried is made to classify standard colon cancer microarray dataset using Association rule mining algorithm, namely Apriori-Hybrid. Apriori-Hybrid, it is the grouping of algorithm Apriori and Apriori-TID, which can classify the large itemsets and can progress the accurateness of classification of cancer and it can also shed light on the basic methodology that enable each cancer type to live and bloom, which in turn help in early recognition of the type of cancer. They propose Apriori-Hybrid as a spontaneous algorithm for tumor classification.

D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu, [3] has presents an effective presentation of the Apriori-Hybrid algorithm over Apriori. Association Rule Mining is a data mining method which is well suitable for mining Market basket dataset. The investigation defined in the previous paper is to prove the possibility of fast accessible data mining algorithms. Although a limited algorithms for mining association rules be present at the time, the Apriori and Apriori TID algorithms significantly compact the overhead costs associated with generating association rules.

R. Agrawal and R. Srikant [4] has presents large database of customer transactions. Each transaction contains of items acquired by a customer in a stay. They present a well-organized algorithm that generates all important association rules between items in the database. The algorithm integrates buffer management and novel approximation and cropping techniques. They also present results of smearing this algorithm to sales data gained from a large retailing company, which shows the usefulness of the algorithm.

Rakesh Agrawal Tomasz ImielinskiArun Swami [5] has describes the difficulty of mining generalized association rules. Given a large database of relations, where each relation consists of a set of items, and a classification (is-a hierarchy) on the items, we find associations between items at any level of the taxonomy. For example, given a taxonomy that says that jackets is a outerwear is clothes, we may infer a rule that “people who buy outerwear tend to buy shoes”. This rule may hold even if procedures that “people who buy jackets tend to buy shoes”, and “people who buy clothes to buy shoes” do not hold. However, this “Basic” algorithm is not very fast; they were existing two algorithms, Cumulate and Merge, which run 2 to 5 times quicker than Basic (and more than 100 times faster on one real-life dataset).

ZijianZheng, Ron Kohavi, and Llew Mason [6] has presents as an Association Rule mining and suggests an algorithm that associated the simple association rules resulting from basic Apriori Algorithm with the multiple minimum maintenance using maximum constraints. The algorithm is employed, and is compared to its ancestor algorithms using a novel proposed comparison algorithm. Results of applying the planned algorithm show faster performance than other algorithms without scratching the accurateness.

Jiawei Han, Jian Pei, Yiwen Yin and Runying Mao[7] has propose new Frequent-Pattern tree(FP tree) structure which is stretched prefix-tree structure for storing compacted, critical information about frequent patterns and established an effective FP tree based mining approach.

FionnMurtagh Mohsen Farid [8] has describes Rare association rules are those that only seem rarely even though they are highly associated with very precise data. In significance, these rules can be very suitable for using with educational datasets since they are usually imbalanced. They explore the mining of rare association rules when congregation student usage data from a Model system. This type of rule is more difficult to find when applying out dated data mining algorithms. Thus they shows some applicable results achieved when comparing numerous frequent and rare association rule mining algorithms. They also offer some descriptive examples of the rules exposed in order to determine both their performance and their worth in educational environments.

3. PROPOSED METHOD

From above literature review we studied previous work done by many people on association rule mining on huge quantity of datasets explains its importance in real life. Now by considering their work and by keeping in mind all the ideas they has developed regarding association rule we are going to proposes association rule mining using APRIORI-hybrid algorithm. After applying Apriori-Hybrid algorithm we proceed the outputted result for comparison purpose to the next algorithm of association rule that is FDM (Fast distributed mining) to compare efficiency, frequency, memory consumption and other parameters. As we know the data stored in a text database is semi-structured that means it is not completely structured as there in a normal database data. And not completely unstructured means a complete huge continues related data.

Consider a data of book store in that the data is completely structured as in the form of title of book, author, publication, year, and cost. And unstructured data in form of abstract and content of books. So mining such semi-structured data for pattern analysing for depth knowledge of the pattern itemsets is nearly difficult to search in todays required huge datasets. So we are proposing in this paper Apriori-Hybrid algorithm. First we parse a huge datasets by our java parser to extract required itemsets from a datasets. Then apply Apriori-hybrid algorithm for association mining. And then we compare accuracy, efficiency and other parameters by hybrid algorithm and FDM (Fast distribution mining) algorithm and represent it graphically to show the result of our output on comparison with other association mining methods.

The main aim of proposed work is to extract interesting patterns from very large text corpus for the purposes of discovering knowledge. It is an interdisciplinary field involving Information Retrieval, Text Understanding, Information Extraction, Clustering, Categorization, Topic Tracking, Concept Linkage, Computational Linguistics, Visualization, Database Technology, Machine Learning, and Data Mining.

APRIORI_HYBRID Algorithm:

Apriori-Hybrid algorithm is a form of algorithm where there is a combination of two apriori algorithms are used. Such algorithms are Weighted Apriori Algorithm and AprioriHashT algorithm. The benefit of both the algorithm is not avoided, so that a new algorithm is invented with the positives points of both the algorithms weighted apriori and THashapriori algorithms. To overcome the negatives points or drawbacks and to combine the positive methodologies of both the algorithms of previous work and generate new tree-Based hybrid method is being introduced in the proposed work.

Flow of Proposed working:

Fig.1. describe the proper working of the process in actual sequence. The process will start by uploading dataset in java**.** To apply parse technique on dataset to extract the required itemsets. It is proposed to find out different kinds of interesting patterns from a set of data with item weight or transaction weight. The weights in these approaches may be thought of as an delay of traditional support in association -rule mining. Weighted association rules can be discovered in a variety of forms, like weighted association rules, fuzzy weighted association rules, and weighted utility association rules.

Then apply aprori hybrid algorithm on dataset. To increase the performance and accuracy of apriori algorithm it is using the hashing data structure. Weighted association -rule mining is concerned with the analysis of consequence of items or transactions in a set of data.

Hash based Apriori implementation, uses a data structure that directly represents a hash table. This algorithm proposes overcoming some of the weaknesses or drawbacks of the apriori algorithm by reducing the number of candidate k-item sets. In particular the 2-itemsets, since that is the key to improving performance. This algorithm uses a hash based technique to reduce the number of candidate item sets generated in the first pass. It is claimed that the number of item sets in C2 generated using hashing can be small so that the scan required to determine L2 is more efficient.

For example, when scanning each transaction in the database to generate the frequent 1-itemsets,L1, from the candidate 1-itemsets in C1, It can generate all of the 2-itemsets for each transaction, hash(i.e.) map them into the different buckets of a hash table structure, and increase the corresponding bucket counts . A 2-itemset whose corresponding bucket count in the hash table is below the support threshold cannot be frequent and thus should be removed from the candidate set. Such a hash based apriori may substantially reduce the number of the candidate k-item sets examined.

In this algorithm, each transaction counting all the 1-itemsets. At the same time all the possible 2-itemsets in the current transaction are hashed to a hash table. It uses a hash table to reduce the number if candidate item sets. When the support count is established the algorithm determines the frequent item sets. It generates the candidate item sets as like the weighted Apriori algorithm**.**

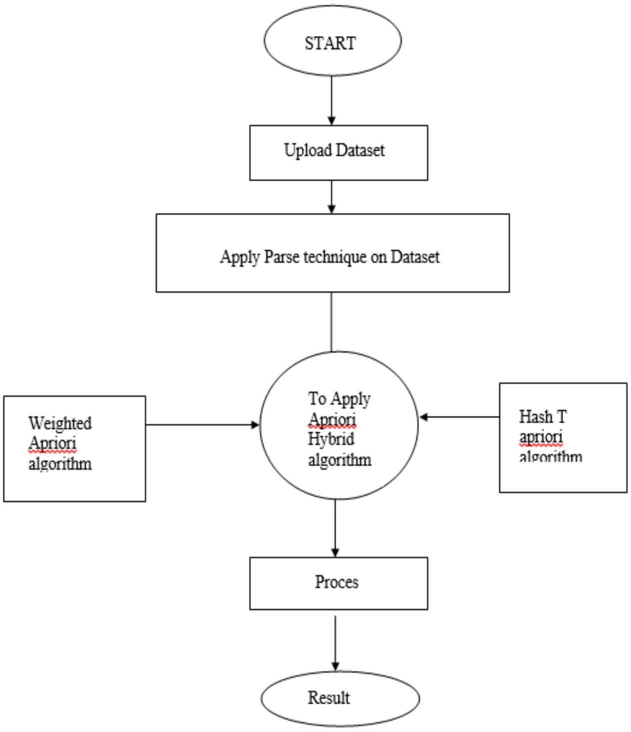

Fig.1. Flowchart of working of system

4. CONCLUSIONS

The main aim of proposed work is to extract interesting patterns from very large text corpus or huge datasets for the purposes of discovering knowledge. It is an interdisciplinary field involving information retrieval, text understanding, information extraction, clustering, categorization, topic tracking, concept linkage, computational linguistics, visualization, database technology, machine learning, and data mining.It will be extracting the required data to analyze via the data parsers available in java. Apriori hybrid algorithm Approach algorithm will be applied. It will process the dataset by the combination of weighted apriori and hash tree algorithm. Efficiency, frequency, memory consumption and other parameters will be check for an algorithm. It can represent graphical representation of the comparison of apriori hybrid with any other mining frequent association algorithm like FDM. Caching techniques and implementation will be used for getting the frequently accessed data while searching (Apart from the algorithm to increase the efficiency) .It can save our search then compare result with FDM algorithm.

5. REFERENCES

1. A. TamirTassa ” Secure Mining of Association Rules inHorizontally Distributed Databases ” IEEETransactions on Knowledge and Data Engineering,vol. 26, no. 4, April2014.

2. Merry KP, Rabindra Kumar Singh, Swaroop.S.Kumar,“apriori-hybrid algorithm as a tool for colon cancer microarray data classification.”, International Journal of Engineering Research and Development e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com Volume 4, Issue 7 (November 2012), PP. 53-57

3. D.W.L. Cheung, J. Han, V.T.Y. Ng, A.W.C. Fu, and Y. Fu, “A Fast Distributed Algorithm for Mining Association Rules,” Proc. Fourth Int’l Conf. Parallel and Distributed Information Systems (PDIS), pp. 31-42, 1996.

4. R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Proc 20th Int’l Conf. Very Large Data Bases (VLDB), pp. 487-499, 1994.

5. Rakesh Agrawal Tomasz ImielinskiArun Swami “Mining Association Rules between Sets of Items in Large Databases”, IBM Almaden Research Center 650 Harry Road, San Jose, CA 1995.

6. ZijianZheng, Ron Kohavi, and LlewMason,”Real World Performance of Association Rule Algorithms”, KDD 2001.

7. Jiawei Han, Jian Pei, YiwenYin,”Mining Frequent Patterns without Candidate Generation”, SIGMOD Conference 2000.

8. S.Ghorai,A.Mukherjee and P.K. Dutta,” apriori-hybrid algorithm as a tool for colon cancer microarray data classification”, IEEE/ACM Transactions on Computational Biology and Bioinformatics,vol.8,No.3,May/June 2011.

9. Yukyee Leung and YeungsamHung,”A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification”,IEEE/ACM Transactions On Computational Biology and Bioinformatics, vol. 7, no. 1, January/March 2010.

10. Ben-David, N. Nisan, and B. Pinkas, “FairplayMP - A System for Secure Multi-Party Computation,” Proc. 15th ACM Conf. Computer and Comm. Security (CCS), pp. 257-266, 2008.

11. R. Agrawal and R. Srikant, “Privacy-Preserving Data Mining,” Proc. ACM SIGMOD Conf., pp. 439-450, 2000.

12. M. Bellare, R. Canetti, and H. Krawczyk, “Keying Hash Functions for Message Authentication,” Proc. 16th Ann. Int’l Cryptology Conf. Advances in Cryptology (Crypto), pp. 1-15, 1996.

13. R. Agrawal and R. Srikant,” Fast algorithms for mining association rules in large databases”. In VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases, pages 487--499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

14. D. Beaver, S. Micali, and P. Rogaway, “The Round Complexity of Secure Protocols,” Proc. 22nd Ann. ACM Symp. Theory of Computing (STOC), pp. 503-513, 1990.